

09/629,831

1. (Currently Amended) A method of automatically creating a dictionary for clustering text documents comprising:
- inputting a maximum dictionary size;
 - determining a frequency of each word in each of said documents;
 - creating a Hashtable dictionary of most frequently occurring words in said documents as limited by said maximum dictionary size;
 - determining a frequency of phrases in each of said documents that contain only words in said Hashtable dictionary;
 - adding most frequently occurring phrases to said Hashtable dictionary; and
 - outputting said most frequently occurring words and said most frequently occurring phrases as said dictionary.
2. (Currently Amended) The method in claim 1, wherein said determining a frequency of each word comprises:
- removing punctuation and case from said documents;
 - removing stop words from said document;
 - replacing words in said documents with synonyms;
 - removing duplicate words from said documents;
 - adding remaining words to said Hashtable dictionary as limited by said maximum dictionary size;
 - determining said frequency of each word remaining in said Hashtable dictionary; and
 - removing words below a frequency level from said Hashtable dictionary.
3. (Original) The method in claim 2, further comprising inputting one or more of said stop words, said synonyms, and said frequency level.
4. (Currently Amended) The method in claim 1, wherein said determining a frequency of phrases comprises:

09/629,831

removing punctuation and case from said documents;

removing stop words from said document;

replacing words in said documents with synonyms;

adding said phrases in each of said documents that contain only words in said Hashtable dictionary to said Hashtable dictionary;

determining said frequency of said phrases remaining in said Hashtable dictionary; and
removing phrases below a frequency level from said Hashtable dictionary.

5. (Original) The method in claim 4, further comprising inputting one or more of said stop words, said synonyms, and said frequency level.

6. (Currently Amended) A method of automatically creating a dictionary for clustering text documents comprising:

inputting a maximum dictionary size;

performing a first pass for each of said documents comprising:

determining a frequency of each word in each of said documents; and

creating a Hashtable dictionary of most frequently occurring words in said documents as limited by said maximum dictionary size;

performing a second pass for each of said documents comprising:

determining a frequency of phrases in each of said documents that contain only words in said Hashtable dictionary; and

adding most frequently occurring phrases to said Hashtable dictionary; and
outputting said most frequently occurring words and said most frequently occurring phrases as said Hashtable dictionary.

7. (Currently Amended) The method in claim 6, wherein said determining a frequency of each word comprises:

removing punctuation and case from said documents;

09/629,831

removing stop words from said document;
replacing words in said documents with synonyms;
removing duplicate words from said documents;
adding remaining words to said Hashtable dictionary as limited by said maximum dictionary size;

determining said frequency of each word remaining in said Hashtable dictionary; and
removing words below a frequency level from said Hashtable dictionary.

8. (Original) The method in claim 7, further comprising inputting one or more of said stop words, said synonyms, and said frequency level.

9. (Currently Amended) The method in claim 6, wherein said determining a frequency of phrases comprises:

removing punctuation and case from said documents;
removing stop words from said document;
replacing words in said documents with synonyms;
adding said phrases in each of said documents that contain only words in said Hashtable dictionary to said Hashtable dictionary;
determining said frequency of said phrases remaining in said Hashtable dictionary; and
removing phrases below a frequency level from said Hashtable dictionary.

10. (Original) The method in claim 9, further comprising inputting one or more of said stop words, said synonyms, and said frequency level.

11. (Currently Amended) A program storage device readable by machine, tangibly embodying a program of instructions executable by the machine to perform a method of automatically creating a dictionary for clustering text documents, said method comprising:
inputting a maximum dictionary size;

09/629,831

determining a frequency of each word in each of said documents;
creating a Hashtable dictionary of most frequently occurring words in said documents as limited by said maximum dictionary size;
determining a frequency of phrases in each of said documents that contain only words in said Hashtable dictionary;
adding most frequently occurring phrases to said Hashtable dictionary; and
outputting said most frequently occurring words and said most frequently occurring phrases as said dictionary.

12. (Currently Amended) A program storage device as in claim 11, wherein said determining a frequency of each word comprises:

removing punctuation and case from said documents;
removing stop words from said document;
replacing words in said documents with synonyms;
removing duplicate words from said documents;
adding remaining words to said Hashtable dictionary;
determining said frequency of each word remaining in said Hashtable dictionary; and
removing words below a frequency level from said Hashtable dictionary.

13. (Original) A program storage device as in claim 12, further comprising inputting one or more of said stop words, said synonyms, and said frequency level.

14. (Currently Amended) A program storage device as in claim 11, wherein said determining a frequency of phrases comprises:

removing punctuation and case from said documents;
removing stop words from said document;
replacing words in said documents with synonyms;
adding said phrases in each of said documents that contain only words in said Hashtable

09/629,831

dictionary to said ~~Hashtable~~ dictionary:

determining said frequency of said phrases remaining in said ~~Hashtable~~ dictionary; and
removing phrases below a frequency level from said ~~Hashtable~~ dictionary.

15. (Original) A program storage device as in claim 14, further comprising inputting said stop words.
16. (Original) A program storage device as in claim 14, further comprising inputting said synonyms.
17. (Original) A program storage device as in claim 14, further comprising inputting said frequency level.
-